One week left to renew!

Renew before 31 July and stay connected with CFA UK

RENEW NOW →

From Zero to Hero - A Data Scientist's Guide to Hardware Part 1

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

Cookies Settings

Manage cookies

Accept All Cookies

provide practical and technical advice for those data scientists shopping around for a new local desktop or laptop as well as for those seeking to optimising their computational workflows with cloud computing.

This article zeroes in on local hardware resources, catering to readers at all levels of familiarity with hardware specifications. It begins by explaining fundamental concepts, particularly aimed at data scientists who may find technical specifications challenging to grasp. It then proceeds to outline the minimum and recommended specifications for local hardware machines (desktops, laptops, etc.), before delving into more advanced considerations tailored for experienced data scientists with demanding workflows that require leveraging remote cloud-powered resources e.g. cloud GPUs.

1. Decoding Hardware Jargon

Readers with prior knowledge of hardware specifications may opt to skip this section.

For those with less experience in the hardware space, even shopping for a simple laptop can become a rather intimidating experience due to the myriad of different buzzwords and computer-science-specific jargon. Complicating matters further, the hardware industry evolves rapidly. Product offerings change dramatically every few years as new advancements are brought to market.

Let's begin by introducing fundamental concepts that can be particularly helpful, especially for newcomers to the fields of data science and computer hardware:

CPU (Central Processing Unit) is the brain of the computer i.e. the primary component of a computer responsible for executing instructions and performing calculations with some of its key components described as follows:

By clicking "Accept All Cookies", you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

(RAM). CPUs typically have multiple levels of cache, including L1, L2, and L3 caches. L1 cache is the smallest but fastest, while L3 cache is larger but slower. A well-balanced cache hierarchy ensures that the CPU can efficiently access data at different levels of granularity.

- Clock Speed: generates a series of electronic pulses that synchronize the operations of the CPU's components. The clock speed, measured in Hertz (Hz) or gigahertz (GHz), determines how many instructions the CPU can execute per second.
- GPU (Graphics Processing Unit) is a specialized electronic circuit component that accelerates computationally intensive tasks. Unlike CPUs, GPUs excel at parallel processing, enabling faster training of models, processing of large datasets, and running complex algorithms. GPUs are primarily used in computers to handle tasks that would otherwise burden the central processing unit (CPU). By offloading these tasks to a dedicated GPU, computers can achieve better performance and efficiency, particularly in DL (Deep Learning) tasks.
- Cores refer to the individual processing units within the CPU chip that can execute instructions independently. Each core operates as a separate processing unit, allowing the CPU to perform multiple tasks simultaneously with some key features being:
- Parallel Processing: CPU cores enable parallel processing by allowing multiple tasks or instructions to be executed simultaneously. This can lead to improved performance and efficiency, especially in multitasking

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

- Threads: Each core can typically handle multiple threads of execution, allowing for further parallelism within a single core. Hyper-Threading (patented by Intel) or Simultaneous Multithreading (SMT) (AMD equivalent) technologies enable each core to execute multiple threads simultaneously, effectively doubling the number of virtual cores.
- CPU cores vs GPU cores: CPU cores are optimised for general-purpose computing tasks, offering high-speed processing for sequential tasks with low latency. In contrast, GPU cores are specialised for parallel processing, featuring many smaller cores capable of executing numerous tasks simultaneously. While CPUs offer flexibility and versatility, GPUs excel at highly parallelisable tasks such as rendering graphics and executing machine learning algorithms.

RAM (Random Access Memory) is a type of volatile “non-permanent” computer memory (data lost when computer is powered off) that stores data and instructions temporarily while a computer is running. RAM is used by the CPU to access and manipulate data quickly. The capacity of RAM determines how much data and program instructions can be stored simultaneously, and it’s typically measured in gigabytes (GB).

HDD (Hard Disk Drive) stores data permanently (data kept when computer is powered off) on spinning magnetic disks (platters) coated with a magnetic material. They use read and write heads to access and modify data on the magnetic disks, which rotate at high speeds.

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

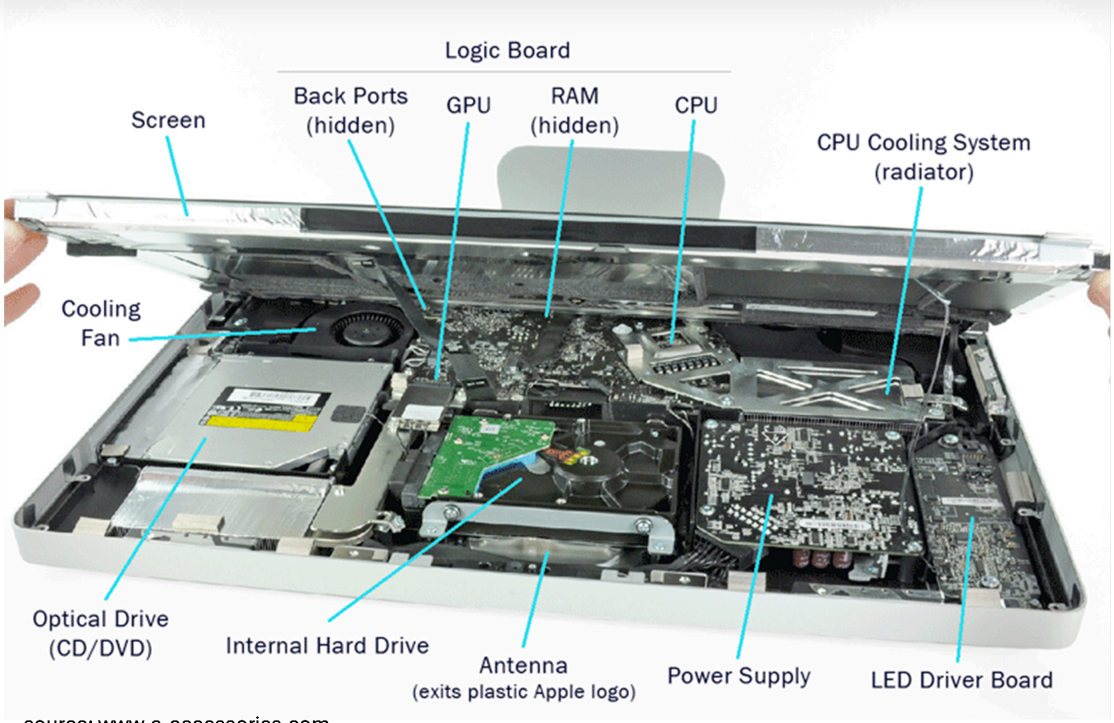
and graphical applications, it can also impact certain data science tasks, especially those involving visualisation, image processing, and DL. GPUs with large VRAM will be able to handle datasets far more efficiently during a DL training workflow.

Graphical Display Resolution is the width and height dimension of an electronic visual display device, measured in pixels, the information from which is used for electronic devices such as computer/laptop monitors:

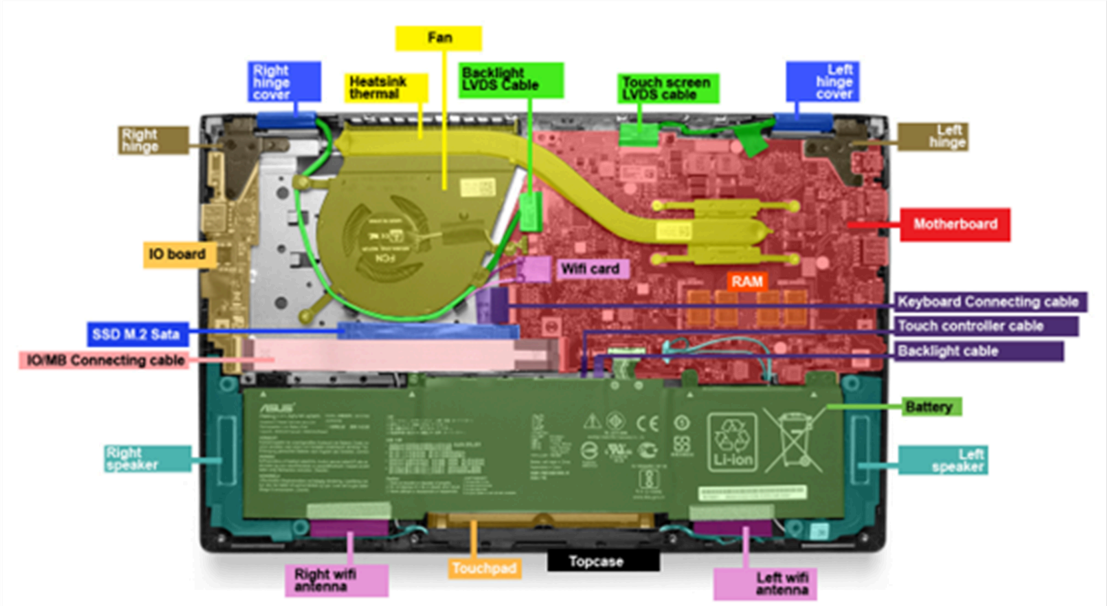
- FHD+ (Full High Definition Plus) typically refers to a display resolution of 1920x1080 pixels that offers a pixel density that is considered high-definition. This resolution is commonly found in laptops, smartphones, and tablets that offers a good price-quality balance.
- UHD+ (Ultra High Definition Plus) is related to a display resolution higher than FHD+ - with a minimum resolution of 3840x2160 pixels - also known as 4K or Ultra HD. UHD+ displays offer significantly higher pixel density and image clarity compared to FHD+, they are also more expensive.
- QHD (Quad High Definition) offers a higher pixel density than FHD+ with 2560 x 1440 pixels, yet its power consumption and price rag fall between FHD+ and UHD+

CFA Society United Kingdom

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.



source: www.a-accessories.com



source: Professional Diploma In Computer Application - PDICA

2. We Need to Talk About GPUs

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

Discrete GPUs were created to tackle this predicament since they are separate components from the CPU and have their own dedicated graphics memory (VRAM). In this way, Discrete GPUs offer higher performance, and they are capable of handling more demanding graphics tasks and data science workflows related to Deep Learning training. For this reason, Discrete GPUs are less affordable than casual integrated GPUs with some examples being the NVIDIA RTX series, or AMD Radeon RX series. A discrete GPU is typically installed in one of the PCIe (Peripheral Component Interconnect Express) slots on the motherboard, which are usually located near the top of the motherboard.

A notable deviation from the previously mentioned details regarding integrated and discrete GPUs lies in Apple's laptops. The Cupertino-based company's approach involves integrating GPU capabilities into its custom-designed system-on-chip (SoC), exemplified by the M chip series, in contrast to the overall industry trend of relying on non-integrated discrete GPUs. This approach reflects Apple's broader strategy, emphasizing vertical integration, performance optimisation, and power efficiency.

Now that a clear distinction between integrated and discrete GPUs has been made, it’s time to discuss the importance of the latter for Data Science tasks. In this way, discrete GPUs are specialised microprocessors crafted for specific tasks, enabling parallel task execution. The primary rationale for leveraging a GPU lies in its advantages for DL (Deep Learning) computations:

- A typical CPU is designed to balance several varying workloads, while a GPU is optimised to do very many simple computations in parallel.
- Offloading tasks from the CPU allows for more concurrent jobs on the same instance, thus alleviating network load.
- The rule of thumb is that GPUs deliver DL output approximately 3x-12x faster than CPUs of equivalent cost for DL tasks depending on the

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

- GPUs are not optimal for serialised CPU tasks such as I/O (input/output) operations, preprocessing, post-processing, and evaluating metrics.
- GPU memory limitations may arise with larger datasets, necessitating multi-GPU configurations, hybrid CPU-GPU setups or more expensive GPUs with higher VRAM.
- GPUs are more expensive than CPUs, especially high-end models required for intensive computations.
- GPUs typically consume more power compared to CPUs, leading to higher energy costs and potentially requiring more robust cooling solutions.

An important fact to consider for data scientists is that acquiring a computer with direct GPU does not automatically qualify the device to perform DL tasks. This is because the most used HPC (High-Performance Computing) frameworks such as CUDA allow machine-learning code to use GPUs, but only by using specific direct Nvidia GPU models. Although other HPC frameworks such as OpenCL are more vendor-neutral than CUDA, it's less user-friendly and less optimised for specific architectures.

3. A Quick Laptop Shopping Checklist for Data Scientists

Now, armed with this knowledge, you should feel more prepared to make informed decisions when selecting a desktop or laptop tailored for data science tasks.

By clicking "Accept All Cookies", you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

- Data science tasks, such as big data handling and ML (Machine Learning), demand a strong processor.
- Choose robust processors like Intel Core i7 or i9, AMD Ryzen 9, M1/M2 (Apple) or higher for efficient task handling.
- An octa-core (8 cores) CPU architecture is the minimum requirement for data science tasks. That said, this choice is influenced by other considerations such as GPU selection. On the one hand, prioritise a higher number of CPU cores over a powerful discrete GPU for low-level tasks (e.g. traditional statistical analysis, small-scale vanilla ML). On the other hand, for those users focused on DL frameworks, consider a configuration with fewer CPU cores but a more powerful discrete GPU.
- For cache memory, aim for at least 4MB to handle basic tasks and small to medium-sized datasets effectively; yet realistically speaking at least 8MB is required for more intensive data analysis and ML tasks. That said, a minimum of 16 MB is advised for handling large datasets and complex DL models efficiently.

RAM

- A reasonable RAM ensures smoother performance and prevents slowdowns during heavy workloads.
- Data scientists need to handle multiple tasks simultaneously, for which reason a minimum of 16 GB of RAM is required, while 32 GB is the

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

durability, and less noise than HDDs.

- Additionally, SSDs are more resistant to physical shocks and vibrations, making them suitable for portable devices like laptops that may be used in various environments.
- A minimum of 500 GB is required, while 1 TB is the recommended SSD capacity.
- Nowadays storage is not the critical feature when acquiring hardware, since cloud and cold external storage options are feasible and reasonably affordable.

GPUs

- Discrete GPUs are a must-have for power-intensive tasks such as ML model training.
- Check if discrete GPUs are compatible with CUDA if DL (Deep Learning) is essential in the data scientist workflow.
- GPUs with CUDA cores and Tensor cores, such as those found in Nvidia's RTX series, are specifically designed for accelerating ML workloads.
- Prioritise GPUs with high-end VRAM with at least 12 GB since DL models with high-dimensional data or many parameters may require significant amounts of VRAM to store intermediate computations during training. Insufficient VRAM can lead to performance bottlenecks or

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

- Clear and detailed displays are essential for data scientists. The recommended laptop screen size is between 15 and 17 inches as they deliver an appropriate balance between performance, health risk, and value received based on the expenditure.
- Laptops with displays above 17 or below 15 inches tend to be less balanced between the price paid and the value obtained. For those seeking larger screens, external displays can help mitigate costs.
- With regards resolution, the recommendation is 1080p (FHD+) or 1440p (QHD) when dealing with a display size of 15 inches.
- Beware high resolutions such as UHD+/4K for coding activities can result in optical health issues and, ultimately, lower productivity. Only those users engaging in more demanding visual pursuits (gaming, 4K movie streaming) should opt for UHD+/4K.
- Avoid OLED (Organic Light-emitting Diode) technology, despite its superior colour quality, as it comes with several drawbacks: shorter battery life, higher cost, and inferior durability i.e. burn-in risk over time.
- Connectivity Options:
- Most of the laptop or desktop offerings come with a reasonable variety of connectivity ports nowadays.
- Ensure you have multiple USB ports (preferably USB 3.0 or higher) to facilitate easy device connections as well as HDMI or DisplayPort for external displays.
- Ethernet ports for reliable internet access and SD card readers for data

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

- Ensure decent battery life (6–8 hours) for on-the-go work without constant power concerns.
- Balance power and portability based on your needs, considering that powerful components and larger displays consume more battery.

The next table summarises the key takeaways of the previous comments on hardware specification:

Spec	Minimum	Recommended
CPU	i7, Ryzen 7, M1	+i9, M2, Ryzen 9
CPU Cores	8	24
CPU Cache	4 MB	16 MB
Clock Speed	4.6 GHz	+5.4 GHz
RAM	16 GB	+32GB
GPUs	Discrete	Discrete & CUDA compat*
GPUs VRAM	8 GB RAM	+12 GB VRAM
STORAGE	SSD, 512 GB	SSD, +1TB
DISPLAY	FHD+, QHD	FHD+, QHD
BATTERY	5 h	7 h

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

data scientists in charge of training large models. Remote GPUs will be the main point of interest in the next sections.

4. Beyond the Local Workstations: Remote GPUs

Local hardware devices, such as desktops and laptops, serve as convenient tools for data scientists when performing basic data science tasks such as traditional ML or EDA (Exploratory Data Analysis). However, the constraints of local hardware become evident when attempting to execute more power-demanding tasks such as DL (Deep Learning) workflows. Employing a laptop or desktop for deep learning purposes is viable only under the following conditions:

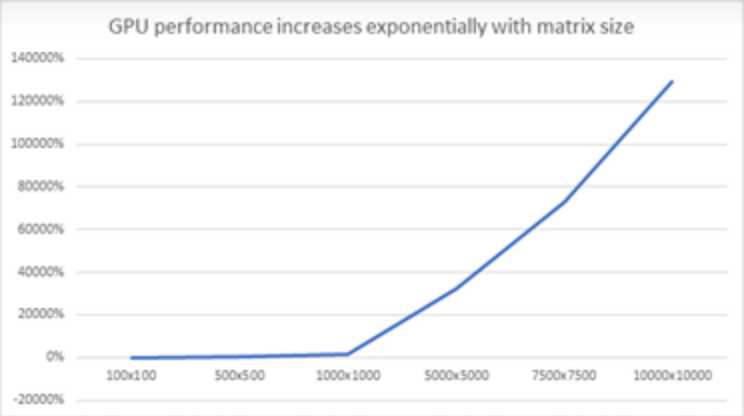
- **Exploratory Work**: For initial exploratory work or prototyping, running DL model training on a local laptop allows for quick experimentation without the need for external resources.
- **Small Datasets**: When working with small datasets that can comfortably fit into the laptop's memory, training locally can be efficient and convenient.
- **Limited Resources**: In cases where access to cloud computing resources or high-performance GPUs is restricted or unavailable, utilising a local laptop for training becomes necessary.
- **Privacy and Security**: When handling sensitive data or proprietary models, training locally ensures data privacy and security without relying on external servers or cloud platforms.

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

That said, DL starts to become insurmountable if any of the last conditions are breached. In such cases, leveraging cloud computing resources may be more appropriate: enter remote/cloud GPUs. There are several reasons why cloud GPUs are often considered a better option:

- **Access to Up-to-Date Hardware**: local machines, especially laptops, might lack powerful GPUs or have outdated hardware, leading to significantly slower training times. Cloud providers regularly update their GPU infrastructure, ensuring access to the latest and most powerful hardware.
- **Scalability**: Cloud platforms offer scalable GPU resources, allowing users to adapt to varying workloads by easily provisioning additional GPU power as needed when dealing with large datasets or complex models.
- **Cost-Efficiency**: Investing in a physical machine with a high-end GPU can be expensive. For instance, an Nvidia Tesla A100 GPU 80 GB is priced around $15,000 to $20,000. To avoid this upfront investment, cloud services provide a pay-as-you-go model, allowing users to pay only for the resources they use, making it more cost-effective for non-regular workloads.
- **Parallel Processing**: Many DL frameworks and libraries are optimised for parallel processing (e.g. torch), a strength of GPUs. Cloud GPUs, designed for this type of computation, can significantly accelerate training times compared to standard CPUs.

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.



The next table offers a comprehensive list of available cloud GPUs with brief comments on their specifications, top-level purpose, advantages, disadvantages, VRAM options, as well as performance capabilities (source: author):

GPU Type	Specs	Ideal For	Pros	Cons	VRAM	FP32	FP16
T4	Balanced performer, cost-effective	Mid-level ML-driven applications	Cost efficiency	Moderate processing	16GB	8.1	65
L4	Entry-level GPU for basic tasks	Small-scale ML projects	Budget-friendly	Limited processing	24GB	30.3	242
H100	High-end GPU for large-scale tasks	Cutting-edge ML research	Peak performance	Higher Cost compared to A100	80 GB	67	1,979
A100	High-end GPU for large-scale tasks	Cutting-edge ML research	Top-Tier Performance	Higher cost compared to V100	40GB, 80GB	19.5	312, 624
V100	Robust GPU for intensive tasks	High-end ML tasks	Cost-effective	Slightly less powerful than A100	32GB	14	28
P100	Balanced Cost-performance	Medium to large-scale ML projects	Balanced perform...	Less powerful than V100, higher cost than K80	16GB	9.3	18.7

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

Performance capabilities are described in the columns FP32, and FP16, which might deserve a more thorough explanation for readers. The values pointed out in these columns are TFLOPS, or Trillion Floating Point Operations Per Second AKA Teraflops, which is a metric that quantifies a processor's computational performance by measuring the number of floating-point operations it can execute in one second. For instance, a GPU with 10 TFLOPS in FP32 can perform 10 trillion floating-point operations per second at this precision. Let's dive into some more details with regards to FP32, FP16 and Mixed-Precision:

FP32 (Floating Point 32-bit): FP32 refers to a data format where numbers are represented in 32 bits (4 bytes):

- It's the standard floating-point format used in many computations, including deep learning.
- FP32 provides high precision, allowing for accurate representation of decimal numbers but requires more memory and computational resources compared to lower precision formats.
- It's commonly used for training deep neural networks where precision is crucial for convergence and accuracy.

FP16 (Floating Point 16-bit): FP16 refers to a data format where numbers are represented in 16 bits (2 bytes).

- It's a lower precision format compared to FP32, which means it can represent a smaller range of numbers and has reduced precision for decimal values.

By clicking "Accept All Cookies", you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

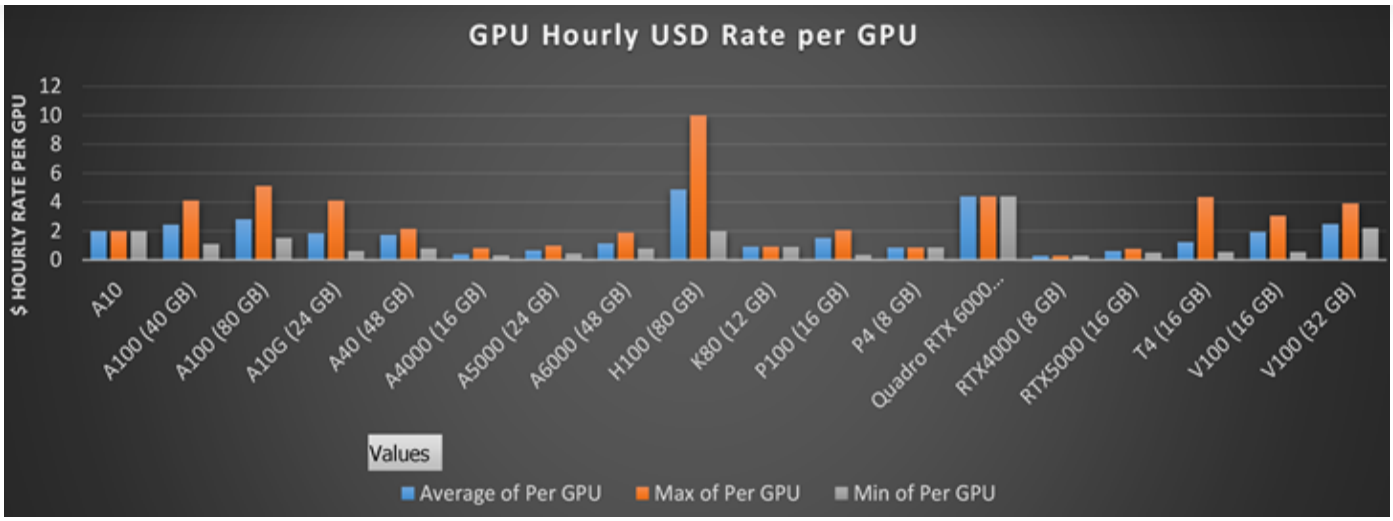
computational efficiency are critical. This includes inference on edge devices or training with large datasets where the reduced precision doesn't significantly impact accuracy.

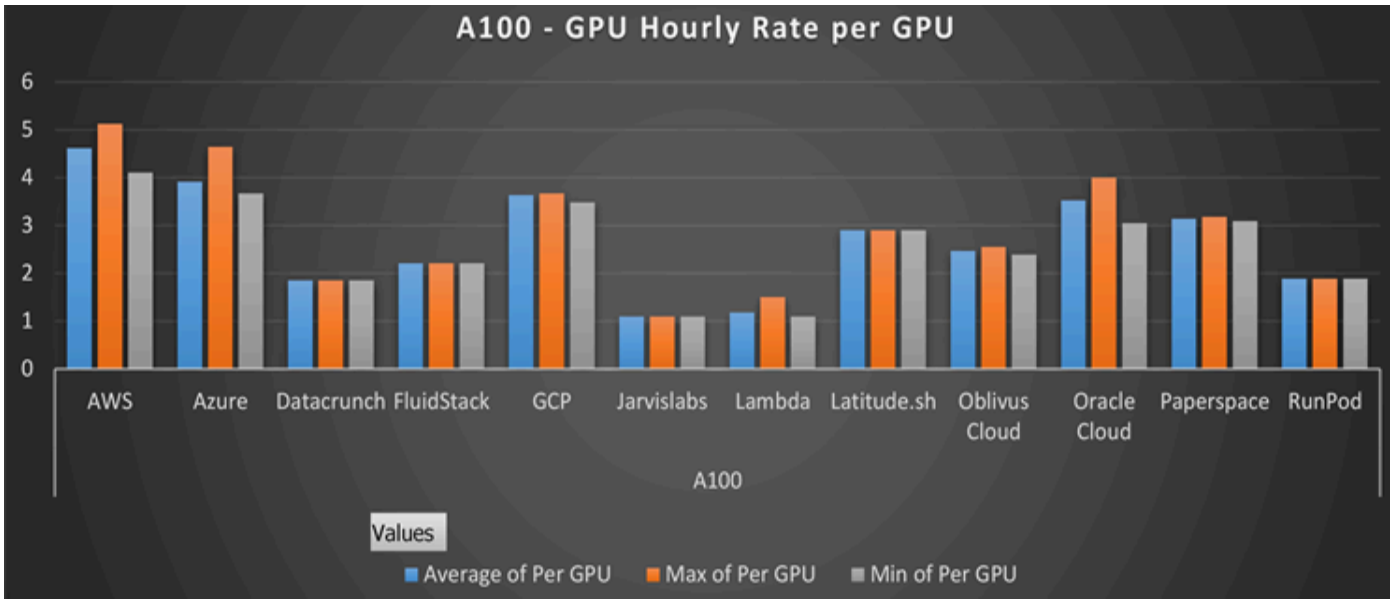
Mixed Precision: although not included in the table, mixed precision is another important metric that refers to the use of both FP32 and FP16 formats within the same computation or neural network:

- In mixed precision training, certain parts of the neural network, particularly those less sensitive to numerical precision, can be computed using FP16. Other parts that require higher precision, such as gradient updates, are computed using FP32.
- This approach leverages the computational efficiency of FP16 where possible while maintaining the necessary precision for accurate model updates and convergence.
- Mixed precision training can significantly accelerate deep learning training workflows by reducing memory usage and computational overhead without sacrificing model accuracy.

To augment the preceding table, the following charts illustrate the average, maximum, and minimum USD hourly rates per GPU sourced from a selection of 15 vendors, encompassing AWS, Azure, GCP, and Lambda, among others (source: fullstackdeeplearning.com). The rates presented are indicative values collected as of December 2023. Before making any decisions, it is advisable to verify the most recent rates from official sources. The initial figure showcases the USD-denominated hourly rate

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.





Although power scalability can be the main factor behind allocating ML workflows in the cloud, there might be other reasons that can make the cloud option the most enticing one such as scalable storage and alternative tools. For instance, cloud providers like Microsoft's Azure can offer built-in tools that can exponentially add value to an organisation or individual. This includes Azure's Cognitive Search retrieval platform which facilitates the search of in-house documents and its combination with LLMs (Large Language Models), which allows you to implement advanced techniques such as RAG

By clicking "Accept All Cookies", you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

substantial computing power or proficiency in data engineering, providing an accessible pathway to enter the realm of cloud-based GPU computing.

Google Colab GPU offers several advantages over using GPU instances on platforms like Google Cloud Platform (GCP), Azure, or AWS:

- **Free Access**: free access to GPU resources, allowing users to execute GPU-accelerated code without incurring additional costs. In contrast, using GPU instances on GCP, Azure, or AWS typically involves paying for computing resources based on usage, which can become expensive, especially for prolonged or resource-intensive tasks.
- **Ease of Use**: Google Colab offers a user-friendly interface and seamless integration with Jupyter notebooks. Users can quickly create, edit, and execute code within a familiar environment, without the need for complex setup or configuration. On the other hand, setting up and managing GPU instances on cloud platforms may require more technical expertise and time.
- **Pre-installed Libraries**: Google Colab comes with a wide range of pre-installed libraries and packages commonly used in data science and machine learning workflows, including TensorFlow, PyTorch, and scikit-learn. This simplifies the development process by eliminating the need to manually install and configure dependencies. While cloud platforms also offer various libraries, users may need to set them up themselves.
- **Collaboration and Sharing**: Google Colab enables seamless

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

- **Integration with Google Services**: Google Colab seamlessly integrates with other Google services such as Google Drive, allowing users to store and access their notebooks and datasets directly from their Google accounts. This tight integration enhances workflow efficiency and enables easy access to data stored in Google Cloud Storage or other Google services.
- **Resource Management**: Google Colab automatically manages resource allocation and usage, dynamically adjusting the available resources based on demand. This ensures optimal performance and resource utilisation without the need for manual intervention. In contrast, managing GPU instances on cloud platforms may require users to monitor and adjust resource configurations manually.

Although Google Colab offers a convenient and cost-effective solution for leveraging remote GPU-accelerated computing resources, particularly for individuals and small teams working on data science and machine learning projects, it also comes with some limitations:

- **Limited Session Duration**: free Google Colab sessions have a time limit for runtime, typically around 12 hours, while subscription users can extend this limit to a maximum 24 hours. After this time, the session will automatically disconnect, and any unsaved work will be lost. This limitation can be inconvenient for long-running computations or training sessions.

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

limitations restrictive and, eventually, will have to sign up for a paid subscription to have access to more powerful GPUs.

- **Dependency on Google Services**: Google Colab relies on Google's infrastructure and services. While this provides seamless integration with other Google services like Google Drive, it also means that users are dependent on Google's uptime and service reliability.
- **Limited Offline Functionality**: Google Colab requires an internet connection to access and run notebooks. While it's possible to download notebooks for offline use, the execution of code and interaction with the runtime environment is restricted when offline.
- **Sharing Limitations**: While Google Colab allows for collaboration and sharing of notebooks, there are limitations on the number of users who can access a notebook simultaneously. Additionally, sharing may require users to have Google accounts and permissions may need to be managed carefully.
- **Variable Performance**: The performance of Google Colab can vary depending on factors such as server load and resource availability. Users may experience fluctuations in performance, especially during peak usage times.

Despite these obstacles, Google Colab provides a convenient and economical solution for accessing GPU-accelerated computing resources, especially beneficial for individuals and small teams engaged in data science and machine learning projects. You can freely

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

In the domain of deep learning, choosing the appropriate GPU for training models is pivotal to attain peak performance,. However, to make an informed decision, data scientists must scrutinise GPU benchmarking results. These GPU benchmarking studies evaluate the training duration needed for popular deep learning models on tasks akin to the researcher's objectives and datasets bearing similarities to those intended for use.

Simply put, selecting the optimal GPU model for a text sentiment analysis task based on benchmarking results designed for a computer vision task would be ill-advised. For instance, when comparing H100 and A100 GPUs capabilities for DL (Deep Learning), the researcher should understand the specific study type and model architecture to be implemented before making any decision. This is because H100 can be faster but also a more expensive alternative than a A100 GPU instance: enter GPU benchmarking. In this way, H100 is able to work 4x faster than A100 for NLP/LLMs models; whereas this edge is reduced to approximately 2x faster performance (source: lambdalabs.com).

GPU benchmarking is particularly important to assess the performance and efficiency of GPUs for training DL (deep learning) models, helping users make informed decisions when selecting hardware for their projects. In this way, a standard GPU benchmarking analysis follows the similar generic steps:

- **Selecting Models and Datasets**: Select appropriate DL models and standard datasets that align with the specific task at hand. For example, when developing a sentiment analysis model, it's essential to avoid using models and datasets designed for computer vision tasks, such as VGG-16 and CIFAR-10. Instead, opt for models like BERT and datasets like IMDb, which are tailored for sentiment analysis tasks.
- **Setting Up the Benchmarking Environment**: Install necessary DL frameworks (e.g. TensorFlow, PyTorch, or Keras) along with GPU drivers

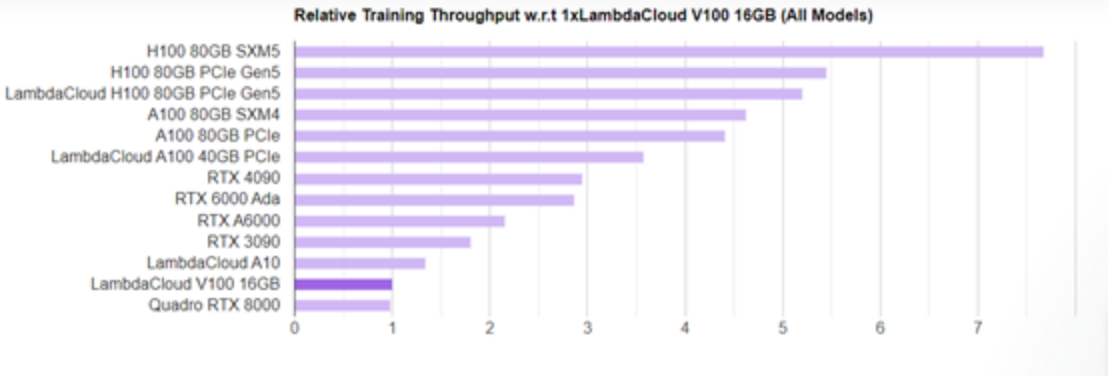
By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

- **Evaluating Benchmarking Results**: Compare the training times across different GPUs by considering factors such as GPU architecture, memory bandwidth, and compute capability; among many others. Identify the GPU that offers the best balance of performance, efficiency, and cost for your specific deep learning task.

Given the costly nature of GPU benchmarking, it is recommended that data scientists refer to regular reports conducted by various entities that evaluate GPUs using established models (e.g. RESNET, VGG-16, etc) along with vanilla datasets (e.g. IMDb, ImageNet, etc).

- Arxiv serves as an excellent starting point, particularly for researchers seeking papers related to GPU benchmarking. Conducting a keyword search for papers titled "GPU Benchmarking" can yield insightful results, depending on the specific machine learning task the researcher intends to undertake e.g. news sentiment analysis for trading.
- Lambda Labs, a Cloud GPU vendor, offers an interactive tool to review GPU benchmark results on a complimentary basis. The sole drawback lies in the limited variety of models available, predominantly focusing on computer vision models. Nevertheless, it's a good starting point for those researchers that wish to perform a preliminary analysis on GPU selection. For instance, the next chart shows the FP32 performance of different GPU models against the base benchmark (V100 16GB) across all the models:

By clicking "Accept All Cookies", you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.



- STAC (Strategic Technology Analysis Center) Benchmark council studies provide rigorous, industry-standard tests designed to assess the performance of hardware and software solutions in financial tasks such as algorithmic trading, backtesting or derivatives pricing among many others. These studies are essential for financial firms to make informed decisions about the selection and optimisation of their technology infrastructure e.g. GPU selection.

Data Scientists using their personal budget to access cloud GPU computing can crunch some numbers by approximating their GPU budget, estimating a range of GPU USD-denominated hourly rates and coming up with a reasonable ballpark figure of training hours per week, month or year. For instance, the next table is calculated using the next parameters:

- $3,000 as annual cloud GPU budget.
- Data scientist has determined that a Nvidia A100 model will be required based on the specific model(s) and dataset(s) to be used for a particular tasks e.g. training NLP model to conduct entity recognition of

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

usage by the hourly cost of the specific GPU. To estimate monthly and weekly usage, the annual allowance is proportionally distributed based on the number of periods within the year e.g. monthly obtained by dividing annual usage by twelve.

	Vendor		
GPU Hours Usage Allowance	Cheap	Mid-Tier	Expensive
Annually	2727	1141	586
Monthly	227	95	49
Weekly	52	22	11

Hence, a data scientist with a $3,000 budget may anticipate their weekly allocation for model training hours to fall within the range of 52 to 11 hours, depending on the chosen vendor. This approximation provides the data scientist with a rough estimate of the expected number of cloud GPU usage hours.

When it comes to corporate GPU budgeting, the procedure becomes more intricate yet even more critical for enhancing performance and controlling expenses in ML projects. Experimentation and iterative refinement remain pivotal in striking the right balance between performance and resource utilisation. In a corporate setting, establishing proper protocols is essential to generate comprehensive documentation. This documentation aids the data science team in justifying the enumeration of a cloud GPU budget, encompassing all the intricacies and specifications involved in the training process.

To illustrate this process, let's use the Tesla A100 GPU as the example for a LSTM model that uses alternative data to predict stock returns and the next assumptions extracted from proprietary experience and third-party insights on similar tasks:

- Training Duration: 50 epochs.
- GPU Type: NVIDIA Tesla A100 (40 GB HBM2 VRAM)

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

- Memory Usage per Batch: Assume 12 GB. Beware this varies depending on the size of the input data and the complexity of the model.
- Total Memory Usage per Epoch: 1,536 GB (12GB per batch for 128 batches)
- Total Memory Usage for 50 Epochs: 76,800 GB (1,536 GB for 50 epochs)

Since the Tesla A100 has 40 GB HBM2 VRAM, we would not be able to fit the entire training process into the GPU memory at once. Therefore, reducing the batch size or the size of the input data might be necessary. Alternative, more advance strategies such as gradient accumulation or model parallelism can be implemented to effectively train the model:

- Gradient accumulation: gradients of the loss function are accumulated over multiple batches before updating the model parameters, which allows for simulating a larger effective batch size without requiring additional memory. That said, this technique may increase training time since weight updates are less frequent.
- Model Parallelism splits the model itself across different GPUs. Each processing unit computes activations and gradients for its assigned part of the model. Communication between processing units is required to synchronise computations:
- Model Parallelism allows for training larger models that may not fit into

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

- Multi-GPU training scales the performance almost linearly with a scale factor of 90-95% (source: AIME).
- The soft spot in model parallelism is that additional GPU instances should be used and, as result, it will result in additional cloud GPU costs.
- Mixed Precision Shift is a feature definitely worth a look in regards of performance is to switch training from float 32 precision to mixed precision training, which can speed-up the training by more than factor 2x (source: AIME).

Overall, TCO (Total Cost Ownership) analysis helps data scientists and their organisations to plan their GPU usage effectively. A more realistic approach is to used professional budgeting software offered by third party vendors or even the same cloud providers:

- Azure Budget Tool
- AWS Budget Tool
- GCP Budget Tool

Conclusions

This article has offered an exhaustive guide for data scientists to navigate hardware choices, including laptops, and desktops. It meticulously covered various aspects such as CPU, RAM, storage capacity, and GPU specifications, with particular emphasis on discrete GPUs. This focus stems from their critical role in deep learning workflows, where

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

process and allows researchers to select resources based on their computational needs and budget constraints.
Sources and Links of Interest:

- CUDA compatible GPUs
- Cloud GPUs costs by vendor and GPU spec
- STAC Benchmark Council Benchmarking Research for Finance Applications
- Lambda Labs GPU Benchmarking Interactive Tool
- Bizon GPU Benchmarking Interactive Tool
- Arxiv GPU Benchmarking Search
- Google Colab Lab Session: Using Cloud GPUs on Google Colab Notebooks

PIC and Links

L/S Portfolio Manager | Lecturer & Consultant

- https://www.linkedin.com/in/csalasls/
- https://twitter.com/CarlosSNaj

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

JUL 2026 » TECHNOLOGY

Can finance keep pace with technological disruption without losing public trust?

FEB 2025 » TECHNOLOGY

The rise of AI and its impact within the investment industry

MAY 2024 » TECHNOLOGY

From Zero to Hero - A Data

NOV 2023 » TECHNOLOGY

A walk through generative AI

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.

Quick links

Media Centre

Privacy Notice

Terms and conditions

Cookie policy

Contact us

Get in touch

3rd Floor, Boston house, 63-64 New Broad street, London, EC2M 1JJ

How to get here

Follow us

By clicking “Accept All Cookies”, you agree to the storing of cookies on your device to enhance site navigation, analyze site usage, and assist in our marketing efforts.